DS3000: Coffee Roast Preference Prediction

Nigel Lobo
nlobo9@uwo.ca

***Abstract*—This paper explores a machine learning approach to predicting coffee roast preferences using data from a survey conducted by YouTuber James Hoffman. 5000 survey participants were mailed 4 coffee samples and performed blind tastings from their homes. They were asked to state their favourite coffee roast, followed by ranking each blind sample and providing data about their demographics, spending, and coffee-related habits. This paper's goal is to predict a participant's stated roast preference (Light, Medium, Dark) using their revealed sensory rankings and habits. The dataset is small and inherently noisy due to the subjectivity and environmental factors that contribute to flavour preferences, so feature engineering and data preprocessing was performed to derive more signal from the data. Four models were trained and compared: Random Forest, CatBoost, XGBoost, and a Multi-layer Perceptron Neural Network. Decision tree models were chosen due to their ability to handle tabular, categorical, and noisy datasets. A neural network was chosen due to its ability to capture non-linear relationships, which is suited for complex, subjective opinions on coffee. Accuracy, Precision, Recall, F-1, and AUC-ROC metrics were reported for comparison. Considering these metrics and the inherent class imbalance, CatBoost performed the best given its F-1 score of 0.6737, which is the harmonic mean of Precision and Recall. However, due to small variations in training runs, any decision tree model could have performed the best, given that the gap between the models is roughly $\pm$ 0.01. The Multi-layer Perceptron performed worse than the decision trees possibly due to the small noisy dataset causing overfitting. Overall, model performance may have been limited by the datasets small size and errors in self-reporting, environmental factors or personal biases. The code and dataset are available at: https://github.com/NigelLobo/coffee-roast-classifier.**

I. Introduction

Popular coffee YouTuber, James Hoffman, conducted a survey of 5000 Americans (4000 respondents) by mailing them 4 coffee samples and having them blindly rank flavour profiles and provide their coffee drinking and spending habits. He uploaded a YouTube video [1] aggregating the results alongside the dataset, but did not apply any machine learning techniques.

In the survey, participants were asked what their favourite roast level was out of Light, Medium and Dark. They then blind tasted and ranked the following coffee samples: Coffee A = Light Roast, Coffee B = Medium Roast, Coffee C = Dark Roast, and Coffee D = Light Roast.

The goal of this project is to predict a participant's preferred coffee roast level (Light/Medium/Dark) from their blind rankings, spending habits and demographics. This is treated as a multi-class classification task. Three decision tree models and a neural network were trained (Random Forest, CatBoost, XGBoost, Multi-layer Perceptron) on the dataset.

As mentioned by Hoffman, 49% of participants stated that they preferred a certain roast before the tasting, but were revealed to prefer a different roast when comparing their blind rankings. It's important to note that sensory ratings are subjective and are affected by environmental factors.

This paper will discuss work related to building machine learning models on coffee datasets, the rationale behind this paper's feature engineering, data preprocessing, methodology and model results.

II. Background

Predicting coffee roast level preferences falls under the category of multi-class classification tasks. In Hoffman's YouTube video, *Surprising And Fascinating Results From The Taste Test* [1], he performs data analysis by showing the distributions of survey questions and coffee rankings. He also discusses the discrepancy between stated roast preference and revealed roast preference, a phenomenon that shows that a respondents believed preference may differ from their true preference. This paper focuses on predicting stated roast preference from survey features using machine learning.

In *Prediction of Coffee Ratings Based On Influential Attributes Using SelectKBest and Optimal Hyperparameters*, the researchers train models to predict coffee quality based on physical measurements, as opposed to subjective sensory ratings as in this paper. The paper suggests that using SelectKBest to identify the optimal set of features and hyperparameters improved the predictive ability of the models and reduced overfitting. This paper will attempt to apply these methods to predicting sensory preferences.

III. Methods

A. System Architecture

The proposed system cleans and derives new features from the raw dataset and SelectKBest (k=20) was used to identify the optimal set of features using ANOVA F-values. Features are imputed and scaled appropriately depending on feature type. Hyperparameters are tuned using GridSearchCV for the decision tree models. K-fold cross validation (k=10) was used for all four models to reduce overfitting. The target variable to be predicted is `stated_roast_level` (Light, Medium, or Dark).

B. Data Preprocessing and Feature Engineering

Every feature of the dataset is categorical due to the subjectivity of the original survey conducted. To derive more signal from the data, new features have been engineered with existing ones being cleaned and converted to numeric data types.

New Features Added:

- `is_purist`: whether or not a person adds milk/sugar to their coffee, derived from the `additions` feature. The idea is that people who drink black coffee may be doing it to appreciate complex flavours otherwise masked by additions
- Brew skill features, derived from `brew` denote whether or not a person uses a particular brewing method based on level of attention and skill:
  - `brew_High_Skill`: person uses Espresso or Pour over
  - `brew_Medium_Skill`: person uses French Press, Bean-to-Cup machines, or Cold Brew
  - `brew_Low_Skill`: person uses Instant Coffee, Pod/capsule machines, or flash-frozen coffee
- `revealed_roast_level`: a mapping from the preferred coffee sample to it's actual roast level. For example, a person who most liked Coffee B would have a revealed roast level of Medium

Modifications to Existing Features:

- `total_spend`, `most_paid`, `most_willing`, `spent_equipment` were converted from dollar ranges to numerical midpoints
- The feature `roast_level` was renamed to `stated_roast_level`, and rare values were re-categorized: Nordic and Blonde roasts were grouped as Light, while French and Italian roasts were grouped as Dark.

Missing values are imputed with the following strategies: (1) Most Frequent for ordinal features, (2) Median for numeric features, (3) Most frequent for nominal features. Numeric features were also scaled using StandardScaler.

TABLE I
SELECTKBEST FEATURES (K=20)

Feature	Description
age	Respondent age demographic
revealed_roast_level	Preferred blind tasting roast
style_Bold	Likes bold notes
style_Bright	Likes bright notes
style_Caramalized	Likes caramelized notes
style_Chocolatey	Likes chocolatey notes
style_Floral	Likes floral notes
style_Fruity	Likes fruity notes
style_Full Bodied	Likes full bodied notes
style_Juicy	Likes juicy notes
style_Nutty	Likes nutty notes
prefer_overall_Coffee A	Most prefers coffee A
prefer_overall_Coffee B	Most prefers coffee B
prefer_overall_Coffee C	Most prefers coffee C
prefer_overall_Coffee D	Most prefers coffee D
favorite_Latte	Favorite coffee style is latte
favorite_Pourover	Favourite coffee style is pour over
know_source_No	Does not know source of coffee brewed
know_source_Yes	Does know source of coffee brewed
gender_Female	Is female
gender_Male	Is male
brew_High_Skill_0	Does not use high skill brew methods
brew_High_Skill_1	Does use high skill brew methods
brew_Low_Skill_0	Does not use low skill brew methods
expertise	1-10 self-rating on coffee expertise

Table I contains the optimal feature set chosen by SelectKBest. All variables are one-hot encoded, except for `age` (ordinal), `revealed_roast_level` (ordinal) and `expertise` (numeric).

C. Algorithm Design

The task is to build a multi-class classification model on a small tabular dataset. Performing a blind tasting survey at home produces inherently noisy data, so careful feature engineering and model selection is important to avoid overfitting on irrelevant patterns. Random Forest and Gradient Boosted Decision Tree models emerged as good starting points due to their handling of categorical, noisy data and their ability to capture non-linear relationships. A neural network was chosen to provide a stark comparison to decision tree models. Four models were chosen for training: **Random Forest**, **CatBoost**, and **XGBoost** and a **Multi-layer Perceptron**. Hyperparameters were tuned for each model during training and Accuracy, Precision, Recall, F-1, and AUC-ROC are reported alongside a confusion matrix for comparison. These metrics were chosen to handle the class imbalance of `stated_roast_level`:

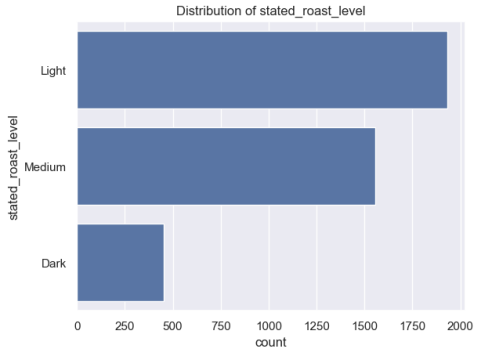

Fig. 1. Distribution of roast preferences.

D. Model Training

The dataset was split 80-20 and `stated_roast_level` was used to stratify the split to maintain the class imbalance. K-fold cross validation was also used with k=10 and shuffling enabled. Below are the tuned hyperparameters/configurations for all four models. GridSearchCV was used to tune the hyperparameters for the decision tree models.

1) Random Forest

TABLE II
RANDOM FOREST HYPERPARAMETERS

Hyperparameter	Value
max_depth	20
max_features	`sqrt`
min_samples_leaf	2
min_samples_split	5
n_estimators	500

This set of hyperparameters strike a balance between model complexity and generalization, potentially reducing overfitting.

2) **CatBoost**

TABLE III
CATBOOST HYPERPARAMETERS

Hyperparameter	Value
bagging_temperature	1
depth	4
iterations	300
l2_leaf_reg	1
learning_rate	0.05

These hyperparameters will allow CatBoost to learn gradually and uses shallow trees alongside Ridge regularization.

3) **XGBoost**

TABLE IV
XGBOOST HYPERPARAMETERS

Hyperparameter	Value
colsample_bytree	1.0
enable_categorical	True
gamma	1
learning_rate	0.1
max_depth	4
n_estimators	100
reg_lambda	5
subsample	1.0

These hyperparameters create a model that is conservative and can generalize well, due to the shallow trees, Ridge regularization and moderate learning rate.

4) **Multi-layer Perceptron**

TABLE V
MLP HYPERPARAMETERS

Hyperparameter	Value
Input Layer	size = 20
Hidden Layer 1	size = 128
Hidden Layer 2	size = 64
Output Layer	size = 3
max_iter	500
optimizer	adam
activation	ReLU

This architecture is wide enough to provide sufficient space to capture non-linear or complex relationships in the dataset.

IV. RESULTS

Below are the performance metrics and confusion matrices for each model after training. The top 15 feature importances are reported for the decision tree models to see which features most contributed to the models prediction.

1) **Random Forest**

TABLE VI
RANDOM FOREST PERFORMANCE METRICS

Metric	10-fold CV	Test Set
Accuracy	0.6796 ± 0.0237	0.6815
Precision	0.6708 ± 0.0345	0.6821
Recall	0.6796 ± 0.0237	0.6815
F1 Score	0.6582 ± 0.0252	0.6653
ROC AUC	0.8089 ± 0.0112	0.7977

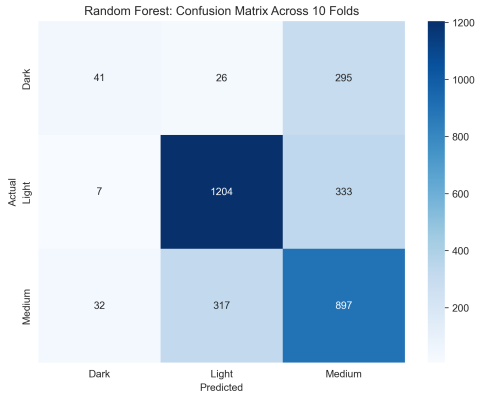

Fig. 2. Random Forest Confusion Matrix

TABLE VII
RANDOM FOREST CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
Dark	0.56	0.15	0.24	91
Light	0.79	0.77	0.78	386
Medium	0.58	0.73	0.65	311

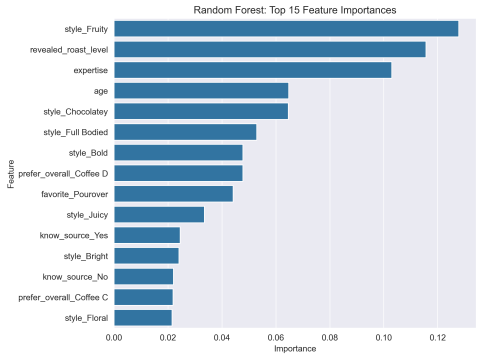

Fig. 3. Random Forest Feature Importances

The Random Forest classifier demonstrated decent results on both cross-validation and the test set. Light roasts were predicted with the greatest accuracy (Precision 0.79 and Recall 0.77), while Dark roasts gave the worst results (Precision 0.56 and Recall 0.15). The most important feature in the prediction was `style_Fruity`, whether or not something liked fruity notes in their coffee.

2) CatBoost

TABLE VIII
CATBOOST PERFORMANCE METRICS

Metric	10-fold CV	Test Set
Accuracy	0.6900 ± 0.0147	0.6891
Precision	0.6873 ± 0.0185	0.6820
Recall	0.6900 ± 0.0147	0.6891
F1 Score	0.6735 ± 0.0182	0.6737
ROC AUC	0.8155 ± 0.0105	0.8057

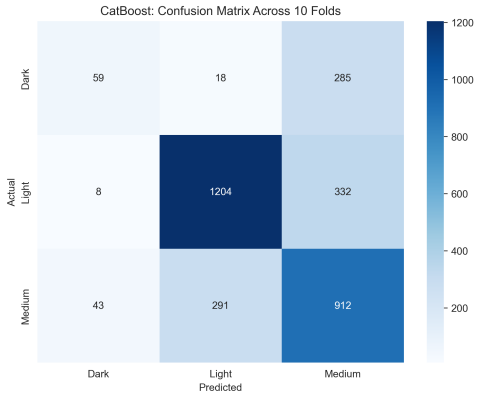

Fig. 4. CatBoost Confusion Matrix

TABLE IX
CATBOOST CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
Dark	0.47	0.15	0.23	91
Light	0.81	0.79	0.80	386
Medium	0.59	0.72	0.65	311

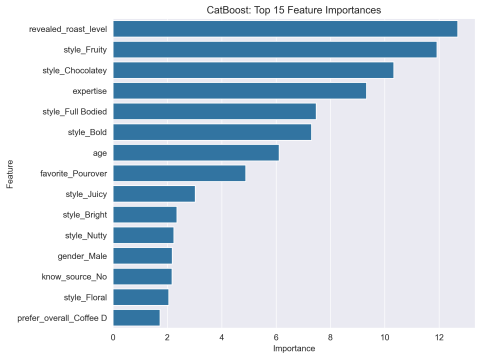

Fig. 5. CatBoost Feature Importances

The CatBoost classifier demonstrated decent results on both cross-validation and the test set. Light roasts were also predicted with the greatest accuracy (Precision 0.81 and Recall 0.79), while Dark roasts gave the worst results (Precision 0.47 and Recall 0.15). The most important feature in the prediction was `revealed_roast_level`, which is the roast value of a respondent's preferred blind sample.

3) XGBoost

TABLE X
PERFORMANCE METRICS

Metric	10-fold CV	Test Set
Accuracy	0.6910 ± 0.0159	0.6878
Precision	0.6870 ± 0.0244	0.6867
Recall	0.6910 ± 0.0159	0.6878
F1 Score	0.6752 ± 0.0188	0.6734
ROC AUC	0.8156 ± 0.0120	0.8070

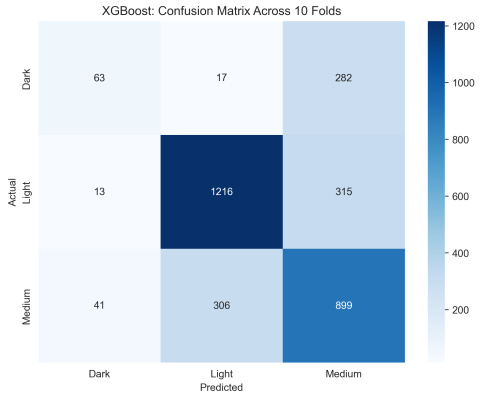

Fig. 6. XGBoost Confusion Matrix

TABLE XI
XGBOOST CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
Dark	0.55	0.18	0.27	91
Light	0.80	0.78	0.79	386
Medium	0.59	0.72	0.65	311

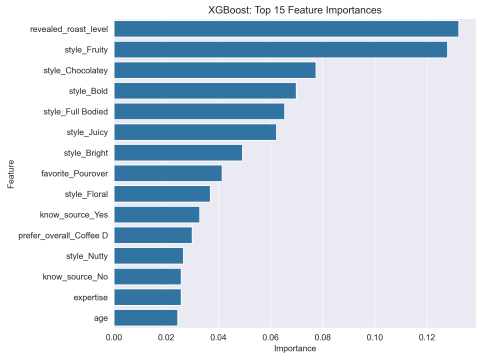

Fig. 7. XGBoost Feature Importances

The XGBoost classifier similarly demonstrated decent results on both cross-validation and the test set. Light roasts were also predicted with the greatest accuracy (Precision 0.80 and Recall 0.78), while Dark roasts gave the worst results (Precision 0.55 and Recall 0.18). The most important feature in the prediction was `revealed_roast_level`, which is the roast value of a respondent's preferred blind sample.

4) Multi-layer Perceptron

TABLE XII
PERFORMANCE METRICS

Metric	10-fold CV	Test Set
Accuracy	0.6136 ± 0.0292	0.6142
Precision	0.6093 ± 0.0280	0.6088
Recall	0.6136 ± 0.0292	0.6142
F1 Score	0.6092 ± 0.0276	0.6100
ROC AUC	0.7285 ± 0.0175	0.7386

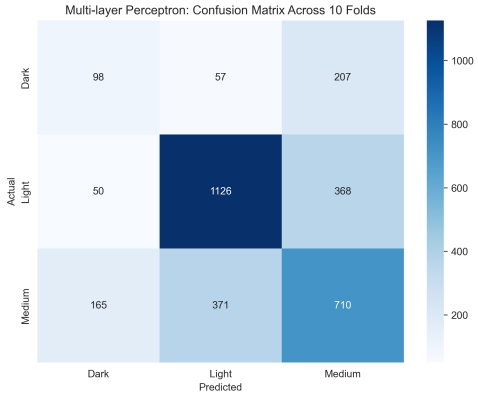

Fig. 8. Multi-layer Perceptron Confusion Matrix

TABLE XIII
MLP CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
Dark	0.23	0.16	0.19	91
Light	0.76	0.73	0.75	386
Medium	0.53	0.60	0.56	311

The Multi-layer Perceptron demonstrated worse results on both cross-validation and the test set, than the decision tree models. Light roasts were predicted with the greatest accuracy (Precision 0.76 and Recall 0.73), while Dark roasts gave the worst results (Precision 0.23 and Recall 0.16). Feature importances can not be seen for neural networks as they act as a black-box because neuron computations are distributed and non-linear.

V. CONCLUSION

Here are the combined performance metrics for all four models trained:

TABLE XIV
TEST SET PERFORMANCE OF ALL CLASSIFIERS

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Forest	0.6815	0.6821	0.6815	0.6653	0.7977
CatBoost	0.6891	0.6820	0.6891	0.6737	0.8057
XGBoost	0.6878	0.6867	0.6878	0.6734	0.8070
MLP NN	0.6142	0.6088	0.6142	0.6100	0.7386

Analyzing the results, the three decision tree models were all very close in performance across the metrics, while the neural network lagged behind. This may due to the datasets small size, inherent noise and subjectivity. To identify the best model, F-1 Score is the ideal metric for handling the class imbalance in `stated_roast_level` because it is the harmonic mean between Precision and Recall. Therefore, **CatBoost** emerged as the highest performer. Unsurprisingly, `revealed_roast_level` was its most important feature. In addition, the respondents feelings about fruity, chocolatey and bold notes were good predictors about their desired roast level. Age and expertise were a toss up, as some models found predictive strength in these features while others did not.

It is also interesting to note that Light roast prediction was significantly better for all models in terms of Precision, Recall and F-1 score than Medium or Dark roasts. The worst prediction occurred with Dark roasts. The class imbalance likely contributed to this difference.

The model results were lower than originally expected but there may be inherent limitations of the dataset due to its small size, and inherent noise in subjective sensory surveys.

Overall, this paper has identified and discussed the most predictive survey features that contribute to a participant's desired coffee roast level and builds on the fascinating results of James Hoffman's taste test.

REFERENCES

[1] J. Hoffman, "Surprising And Fascinating Results From The Taste Test," YouTube, https://www.youtube.com/watch?v=bMOOQfeloH0 (accessed Nov. 13, 2025).
[2] E. Agyemang et al., "Prediction of coffee ratings based on influential attributes using selectkbest and optimal hyperparameters," arXiv.org, https://arxiv.org/abs/2509.18124 (accessed Nov. 13, 2025).
[3] U. Haddii, "The Great American Coffee Taste Test Dataset," Kaggle, https://www.kaggle.com/datasets/umerhaddii/the-great-american-coffee-taste-test-dataset (accessed Dec. 6, 2025).